

- Buchanan, E.A. Feigenbaum, J. Lederberg, G. Schroll, and G.L. Sutherland, "The Application of Artificial Intelligence in the Interpretation of Low-Resolution Mass Spectra", *Advances in Mass Spectrometry*, 5, 314 (1971),
- (25) B.G. Buchanan and J. Lederberg, "The Heuristic DENDRAL Program for Explaining Empirical Data". In proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia (1971). (Also Stanford Artificial Intelligence Project Memo No. 141.)
- (26) B.G. Buchanan, E.A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145.)
- (27) Buchanan, B. G., Duffield, A.M., Robertson, A.V., "An Application of Artificial Intelligence to the Interpretation of Mass Spectra", *Mass Spectrometry Techniques and Appliances*, G. W. A. Milne, Ed., John Wiley & Sons, Inc., 1971, p. 121.
- (28) D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A. Yeo, E.A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference VIII. An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", *Journal of the American Chemical Society*, 94, 5962 (1972).
- (29) B.G. Buchanan, E.A. Feigenbaum, and N.S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In *Machine Intelligence* 7, Edinburgh University Press (1972).
- (30) J. Lederberg, "Rapid Calculation of Molecular Formulas from Mass Values". *Journal of Chemical Education*, 49, 613 (1972).
- (31) H. Brown, L. Masinter, and L. Hjelmeland, "Constructive Graph Labeling Using Double Cosets". *Discrete mathematics*, 7, 1 (1974). (Also Computer Science Memo 318, 1972).
- (32) B. G. Buchanan, Review of Hubert Dreyfus' "What Computers Can't Do: A Critique of Artificial Reason", *Computing Reviews* (January, 1973). (Also Stanford Artificial Intelligence Project Memo No. 181)
- (33) D. H. Smith, B. G. Buchanan, R. S. Engelmores, H. Adlercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". *Journal of the American Chemical Society* 95, 6078 (1973).

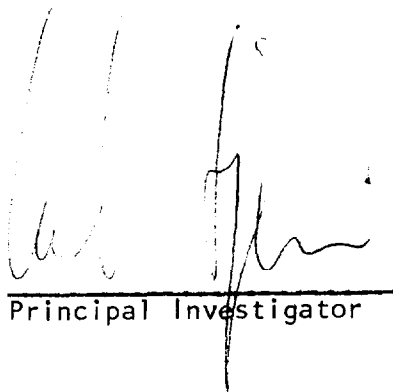
- (34) D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. Intsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". Tetrahedron, 29, 3117 (1973).
- (35) B. G. Buchanan and N. S. Sridharan, "Rule Formation on Non-Homogeneous Classes of Objects". In proceedings of the Third International Joint Conference on Artificial Intelligence (Stanford, California, August, 1973). (Also Stanford Artificial Intelligence Project Memo No. 215.)
- (36) D. Michie and B.G. Buchanan, "Current Status of the Heuristic DENDRAL Program for Applying Artificial Intelligence to the Interpretation of Mass Spectra", in "Computers for Spectroscopy," R.A.G. Carrington, Ed., Adam Hilger, London, 1973. Also: University of Edinburgh, School of Artificial Intelligence, Experimental Programming Report No. 32 (1973).
- (37) H. Brown and L. Masinter, "An Algorithm for the Construction of the Graphs of Organic Molecules", Discrete Mathematics, 8, 227 (1974). (Also Stanford Computer Science Dept. Memo STAN-CS-73-361, May, 1973)
- (38) D.H. Smith, L.M. Masinter and N.S. Sridharan, "Heuristic DENDRAL: Analysis of Molecular Structure," in "Computer Representation and Manipulation of Chemical Information," W.T. Wipke, S. Heller, R. Feldmann and E. Hyde, Eds., John Wiley and Sons, Inc., 1974, p. 287.
- (39) R. Carhart and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XI: The Analysis of C13 NMR Data for Structure Elucidation of Acyclic Amines", Journal of the Chemical Society (Perkin II), 1753 (1973).
- (40) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Application of Artificial Intelligence for Chemical Inference XII: Exhaustive Generation of Cyclic and Acyclic Isomers". Journal of the American Chemical Society, 96, 7702 (1974). (Also Stanford Artificial Intelligence Project Memo No. 216.)
- (41) L. Masinter, N.S. Sridharan, R. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XIII. Labeling of Objects having Symmetry". Journal of the American Chemical Society, 96, 7714 (1974).
- (42) N.S. Sridharan, Computer Generation of Vertex Graphs, Stanford CS Memo STAN-CS-73-381, July, 1973.
- (43) N.S. Sridharan, et.al., A Heuristic Program to Discover Syntheses for Complex Organic Molecules, Stanford CS Memo

- STAN-CS-73-370, June, 1973. (Also Stanford Artificial Intelligence Project Memo No. 205.)
- (44) N.S. Sridharan, Search Strategies for the Task of Organic Chemical Synthesis, Stanford CS Memo STAN-CS-73-391, October, 1973. (Also Stanford Artificial Intelligence Project Memo No. 217.)
- (45) R. G. Dromey, B. G. Buchanan, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra". Journal of Organic Chemistry, 40, 770 (1975).
- (46) D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XV. Constructive Graph Labelling Applied to Chemical Problems. Chlorinated Hydrocarbons". Analytical Chemistry, 47, 1176 (1975).
- (47) R. E. Carhart, D. H. Smith, H. Brown and N. S. Sridharan, "Applications of Artificial Intelligence for Chemical Inference. XVI. Computer Generation of Vertex Graphs and Ring Systems". Journal of Chemical Information and Computer Science, 15, 124 (1975).
- (48) R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure". Journal of the American Chemical Society, 97, 5755 (1975).
- (49) B. G. Buchanan, "Scientific Theory Formation by Computer." In Proceedings of NATO Advanced Study Institute on Computer Oriented Learning Processes, 1974, Bonas, France.
- (50) E. A. Feigenbaum, "Computer Applications: Introductory Remarks," in "Proceedings of Federation of American Societies for Experimental Biology," 33, 2331 (1974).
- (51) R. Davis and J. King, "Overview of Production Systems" To appear in Machine Representation of Knowledge, Proceedings of the NATO ASI Conference, July, 1975. (Also Stanford Artificial Intelligence Project Memo .)
- (52) B. G. Buchanan, "Applications of Artificial Intelligence to Scientific Reasoning." In Proceedings of Second USA-Japan Computer Conference, American Federation of Information Processing Societies Press, August, 1975.
- (53) R. E. Carhart, S. M. Johnson, D. H. Smith, B. G. Buchanan, R. G. Dromey, J. Lederberg, "Networking and a Collaborative Research Community: A Case Study Using the DENDRAL Program," in "Computer Networking and Chemistry", P. Lykos, Ed., American Chemical Society, Washington, D.C., 1975, p. 192.

- (54) D. H. Smith, "The Scope of Structural Isomerism" (Paper XVIII in our series of AI Applications in Chemistry). *Journal of Chemical Information and Computer Science*, 15, 203 (1975).
- (55) D. H. Smith, J. P. Konopelski and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIX. Computer Generation of Ion Structures." *Organic Mass Spectrometry*, 11 (1976) 86.
- (56) B. G. Buchanan, D. H. Smith, W. C. White, R. Gritter, E. A. Feigenbaum, J. Lederberg and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XXII. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program." *Journal of the American Chemical Society*, in press.
- (57) E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green and S. N. Cohen, "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System." *Computers and Biomedical Research* 8, 303-320 (1975).
- (58) R. Davis, B. Buchanan and E. Shortliffe, "Production Rules as a Representation for a Knowledge-Based Consultation Program", accepted for publication by *Artificial Intelligence*. (Also Stanford Artificial Intelligence Project Memo No. AIM-266.)
- (59) R.E. Carhart and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XX. 'Intelligent' Use of Constraints in Computer-Assisted Structure Elucidation," *Computers in Chemistry*, in press.
- (60) C. Cheer, D.H. Smith, C. Djerassi, B. Tursch, J.C. Braekman, and D. Daloze, "Applications of Artificial Intelligence for Chemical Inference. XXI. Chemical Studies of Marine Invertebrates. XVII. The Computer-Assisted Identification of [+-]-Palustrol in the Marine Organism *Cespitularia* sp., aff. *Subvirdis*," *Tetrahedron*, in press.
- (61) T.R. Varkony, R.E. Carhart, and D.H. Smith, "Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in "Computer-Assisted Organic Synthesis," W.T. Wipke, Ed., American Chemical Society, Washington, D.C., in press.
- (62) D.H. Smith and R. E. Carhart, "Structural Isomerism of Mono- and Sesquiterpenoid Skeletons," *Tetrahedron*, in press.
- (63) H. Eggert and C. Djerassi, "The Carbon-13 Magnetic Resonance Spectra of Acyclic Aliphatic Amines," *Journal of American Chemical Society*, 95, 3710 (1973).

- (64) H. Eggert and C. Djerassi, "Carbon-13 Nuclear Magnetic Resonance Spectra of Keto Steroids," *Journal of Organic Chemistry*, 38, 3788 (1973).
- (65) H. Eggert, C. VanAntwerp, N. Bhacca and C. Djerassi, "Carbon-13 Nuclear Magnetic Resonance Spectra of Hydroxy Steroids," *Journal of Organic Chemistry*, 41, 71 (1976).
- (66) S. Hammerum and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXLV. The Electron Impact Induced Fragmentation Reactions of 17-oxygenated Progesterones." *Steroids*, 25, 817 (1975).
- (67) S. Hammerum and C. Djerassi. "Mass Spectrometry in Structural and Stereochemical Problems CCXLIV. The Influence of Substituents and Stereochemistry on the Mass Spectral Fragmentation of Progesterone." *Tetrahedron*, 31, 2391 (1975).
- (68) L. L. Dunham, C. A. Henrick, D. H. Smith, and C. Djerassi, "Mass Spectrometry in Structural and Stereochemical Problems. CCXLVI. Electron Impact Induced Fragmentation of Juvenile Hormone Analogs," *Org. Mass Spectrom.*, in press.
- (69) C. Djerassi, Foreword to "¹³C NMR-Spectroscopy," by E. Breitmayer and W. Voelter, Verlag Chemie GmbH, Weinheim/Bergstr., 1974.
- (70) R. G. Dromey, M. J. Stefik, T. Rindfleisch, and A. M. Duffield, "Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data," *Analytical Chemistry*, in press.

The undersigned agrees to accept responsibility for the scientific and technical conduct of the project and for the provision of required progress reports if a grant is awarded as the result of this application.



Principal Investigator

5/26/76

Date

5 Appendix I

Details of Proposed HELP SYSTEM.

1) On-line documentation system

We designed CONGEN without a highly structured interface between a researcher and the program. This provides a great deal of flexibility in the ways the program can be used to solve a given problem. But this lack of structure can result in a feeling of helplessness when a researcher has little idea of what to do next. The printed document is usually inadequate; it is too long to find a necessary piece of information quickly.

A recent formulation of guidelines for humanizing computerized information systems (T.D. Sterling, Science, 190, (1975), p. 1168), places particular emphasis on the importance of permitting scientists to control interaction with the program. Illustrations of the sorts of help we propose are in the programs MLAB and Interlisp-Masterscope. A demonstration of this concept in the current version of CONGEN is found in the "information interrupts" described previously. In keeping with this user-driven form of interaction, it has become obvious that a flexible, on-line help system in the form of access to the information contained in the document is necessary.

We will rearrange the document into a form that will make it serviceable as a help file, as well as more readable as a reference. Requests for assistance to CONGEN will result in accessing the help file for a summary of what is useful to do at that point, or what commands can be used, or what format is necessary for a given command. Options will be provided for a more detailed description if the researcher finds it necessary for clarification.

2) Tutorial error handling

A new tutorial error handling portion of CGHELP will rely heavily upon a flexible error-detection mechanism in CONGEN, one which is significantly easier to work with than the current version. Errors in CONGEN are now perceived in the traditional manner: built into the code at many different points there are various consistency checks, and when one or more of these is violated, an error message is printed and corrective action is taken (this is usually a simple return to the top-level prompt of the program). This approach has become increasingly more cumbersome as CONGEN has been extended. As each new concept is added to the system all possible conflicts with old concepts must be considered and a progressively larger number of new tests must be added throughout the code. To alleviate these difficulties and to lay the foundation for other CGHELP developments, we plan a completely new approach to the problem of error detection, one which draws upon a knowledge base, external to CONGEN itself, of error conditions. Philosophically, this amounts to a realization that error checking can be dealt with as an activity quite apart from the symbol-manipulation algorithms of the main program.

We intend to formulate this separate error-checking program as a production system which will process each input from the user. In this system, all documented knowledge about CONGEN will be represented internally as a set of situation-action rules. The situation of each rule will be a condition which must not occur during a CONGEN session, and the action portion will be executed whenever the control program detects that a given situation is satisfied. In the first implementation, each action will simply cause an error message to be printed and will provide the "tutor" with information concerning pertinent sections of the document. However, further CGHELP developments (see below) will depend heavily on more sophisticated actions, and in fact the production system will form the core of the "intelligent" aspects of the entire CGHELP system.

The flexibility of the production system format here will allow us to approach the error-detection problem in a general context. The rules themselves must, of course, represent specific knowledge about CONGEN, but the elements of the control system (i.e., the portion of the program which controls testing and evaluation of the rules) will contain the protocols for printing messages and guiding the tutorial interaction. To apply the system to a new program, we will need a new knowledge base and on-line document, but many of the details about interacting with users in a tutorial mode will be directly transferrable.

3) Internal model of the user

In order to accomplish the "tutoring" outlined above without seeming overly solicitous or overly presumptuous, the error handling system will clearly need some model of the user to guide it. For example, an experienced user who types poorly would quickly tire of frequent offers to display the menu of available commands, but to a novice such offers could be quite useful. The user model we plan for CGHELP will contain an internal representation of the user's knowledge of various key concepts in the program, and as he gains new information, either through use of the on-line document or via tutorial error handling, the model will be updated. The tutoring process will then be coupled strongly to this model so that access to the document is offered only when CGHELP perceives the topic as one which has not frequently (or recently) been touched upon.

The coupling of CGHELP to the user model will again be accomplished via the production system concept. The action portion of the error-testing rules mentioned above will be modified so that they cause no direct user interaction. Rather, they will cause internal assertions to be made that an error has occurred. A new body of rules, representing knowledge about how to deal with errors in the context of the user model, will be accessed to generate appropriate actions for the current user. Still other rules, invoked whenever the user accesses the document or otherwise indicates an increased knowledge of the program (say, by flawlessly executing a complex input sequence), will be responsible for updating the model itself.

Ideally, the model for a particular user should span several sessions with the program, and rules should be included which account for the normal attrition of user knowledge over a period of weeks or months. This implies that some profile be stored in the computer system on a long-term basis for each CONGEN user. We will design such a system with care, storing profiles only with the express consent of each user, and will provide alternative methods of defining an initial user profile (e.g., a short question-and-answer period at the start of a session) for those who do not wish to have stored profiles.

4) Error correction

So far we have discussed CGHELP as a system primarily for presenting the user with documented information, allowing him to learn the "rules" of CONGEN as easily as possible. There are of course other functions for a help system and at this point we will begin to explore more general CGHELP tasks.

A frustrating aspect of many interactive programs including CONGEN is that when an error occurs, it is frequently necessary for the user to "back up" and restart the program at some earlier point, even when the error is a relatively simple one which could be corrected locally, at the point of detection. For many user errors in CONGEN it is possible to define one or more probable fixes to the problem. These corrections may be either automatic modifications of internal CONGEN variables or minor digressions from the normal input sequence to allow the user to correct the error or omission himself. The next step in CGHELP development will be to incorporate error correction information into the "actions" of the error detection rules and to establish methods of using this knowledge to help the user recover gracefully from error conditions.

Automatic error correction must be approached with care because it will require CGHELP to take an active role in modifying the user's inputs to CONGEN. This can cause serious difficulties when the presumed correction is not appropriate; blatant errors can be transformed into more subtle ones which are extremely hard to detect later. One of the primary design criteria in the CGHELP error correction system will be that no modification is ever carried out unless CGHELP both obtains an explicit OK from the user and determines, from the user model, that he has an understanding of the nature of the problem and its solution. A second problem is that the automatic correction could take substantially longer than the user himself would need to correct the same problem. In CGHELP we will include some estimation of the lengthiness of possible fixes which, together with a measurement of system load, will influence the selection of the appropriate corrective action.

The natural result of including and maintaining a sophisticated error correction facility in CGHELP will be an increased flexibility in the input language. The user will be

allowed to deviate from the normal input protocols and the burden of verifying the overall correctness of the input will fall upon the program. The messages from the program can be phrased in such a way that the user seldom needs to know that technically he has made "errors" - he will use the commands in an order which seems logical to him and CGHELP will establish the dialog necessary to educate him and to query him as detailed information becomes important.

5) Extensions of error correction to "soft errors"

When we interact with new researchers who are learning to work with CONGEN, we find ourselves explaining not only the "rules" of the program but also many other topics such as strategy, helpful hints, details of the algorithms, etc. The clues that a user needs such higher-level help usually come directly from his inputs to the program, augmented by our mental model of the expectations which users bring to the program. The last phase of CGHELP development will be an open-ended exploration into the automation of help on what we term "soft-errors", or errors which are correct statements but show poor strategy, poor use of commands, and so forth.

We plan several new tools for the error detection system which will allow it to perceive these "soft errors". First, we will develop methods of estimating the computer time and storage space required in specific cases by each of the major functions of CONGEN. Currently there are no guidelines to help a user determine whether a given phrasing of a problem is possible to solve with CONGEN, and cases which are too large cause the program to carry out extensive computations before it becomes obvious to the user that the task is impossible. Second, we will incorporate a scanning facility which can examine intermediate results of a computation as they are being produced, looking for unusual or characteristic chemical features which the chemist may not have realized were possible. The chemical knowledge base which will define the criteria of "unusualness" will be distilled from our experience with typical CONGEN cases, and the chemist will have access to these criteria so that he can change them to better suit his needs, if necessary. Finally, we will create a strategy section which, given a problem in a particular state, will rank, in terms of overall problem efficiency, the possible sequences of commands needed to complete the problem. The evaluation will draw upon the estimator described above and upon a set of heuristics concerning "good form" in approaching CONGEN problems. Such evaluations will give us not only a yardstick against which to measure the user's strategy, but also a possible "driver" for automatically carrying out whole problems.

These tools represent measures of "soft error" conditions which we now feel to be important, but it is likely that other tools will become evident as we gain experience with other users. Perception of "soft errors" will be implemented by adding

appropriate situation-action rules to the basic production system, and the on-line document and tutorial systems will be augmented with information about problem size, chemical unusualness (as defined in CGHELP) and strategy. The user model will gain new importance in this process because it will become an integral part of the decision as to whether or not a "soft error" has even occurred; these conditions are defined in terms of the user's expectations and desires. Also, in order to maintain a considerate and useful dialog between CGHELP and the user, we will explore the inclusion of some elements of user psychology into the model. Because the danger of frustrating or boring the user will be substantially increased when CGHELP takes a more active role in the session, a commensurately more accurate model of user irritation or satisfaction will be needed to guide the program.

6 **Appendix II**
 1975-76 Annual Report to NIH

Table of Contents

Section		Page
	Subsection	
1.	PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE . . .	1
1.1	Introduction	1
1.2	Hardware Acquisition and Development	3
1.3	Software Development	4
1.4	Operating System	6
1.5	Combined Gas Chromatography/High Resolution Mass Spectrometry	7
1.6	High Resolution Spectra Utility Programs	10
1.7	Process Monitor: PMON	11
1.8	METASYS	12
1.9	Summary	13
2.	PART 2: DEVELOPING PERFORMANCE AND THEORY FORMATION PROGRAMS TO ASSIST IN BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS	13
2.1	Introduction	13
2.2	CONGEN	14
2.3	PLANNER	21
2.4	Meta-dendral Rule Formation Programs	21
2.5	Results	26
2.6	Heuristic Programming Project Workshop	27
3.	PART 3: APPLICATIONS TO BIOMEDICAL STRUCTURE ELUCIDATION PROBLEMS	28

3.1	Introduction	28
3.2	Applications by Professor Djerassi's Research Group	29
3.3	Utilization of the Mass Spectrometry Resource	36
3.4	Applications of Programs by External Scientists	38
3.5	Export of GC/MS Programs to Other Sites.	41
	Index	60

II.A. DESCRIPTION OF PROGRESS

OVERVIEW

In the period August, 1975 to July, 1976 the DENDRAL programs and the gas chromatography/mass spectrometry (GC/MS) data system have made significant progress toward the goals stated in the research proposal. This report of progress is organized in three parts, corresponding to the three specific aims of our December, 1973, proposal: (PART 1) Enhancing the power of the mass spectrometry resource, (PART 2) Developing performance and theory formation programs, and (PART 3) Applying the computer programs and instrumentation to biomedically relevant structure elucidation problems.

The DENDRAL project, one of the major users of the SUMEX-AIM computer facility, has been forming its own community of remote users. This "exodendral" community has already provided valuable contributions to program development and both the community and contributions are expected to grow at an increased rate. Our programs are receiving heavy use from local users and outside users who are investigating structure elucidation problems for a variety of different compound classes. Local users include members of Professor Djerassi's group, other chemistry department persons and research groups at the Stanford Medical School. We have continued building a community of outside users who can access our programs at SUMEX through the TYMNET or ARPANET.

1 PART 1: ENHANCING THE POWER OF THE M.S. RESOURCE

1.1 Introduction

Our grant proposal requested funds for significant upgrading of our capabilities in mass spectrometry. The goals of this upgrading were to provide routine high resolution mass spectrometry (HRMS), combined gas chromatography/low resolution mass spectrometry (GC/LRMS) and to develop a combined gas chromatography/high resolution mass spectrometry (GC/HRMS) facility. In addition, this would provide the capability for new experiments in the detection and utilization of data on metastable ions. These capabilities would then be available as required for application to our wider goal, solution of biomedical structure elucidation problems of a community of researchers.

The upgrading included several items of hardware and software development, as follows: 1) Acquire stand-alone computer support for the mass spectrometer because existing facilities were inadequate and very expensive; 2) convert existing software, written in the PL/ACME language into FORTRAN so that it would run on the new system; 3) develop new software as required for the demanding task of GC/HRMS; 4) provide hardware and software for semi-automatic acquisition of data on metastable ions. The initial development phase of this upgrading included performance tests to determine the capabilities and limitations of the GC/HRMS system to define the scope of problems to which it can be applied. The past year's efforts (year two of the grant) have culminated in accomplishment of many of the above goals for development. In the first year, the computer system (a Digital Equipment Corp. PDP 11/45) was purchased, installed and is now operating routinely in conjunction with the mass spectrometer (a Varian-MAT 711) and an auxiliary PDP 11/20 system (see system configuration, Fig. 1). Program conversion and modification for the initial version of the software system was completed and the computer system now provides complete stand-alone support for our experiments in mass spectrometry. Over the past year we have developed further our philosophy of data acquisition and reduction based on computed models of the actual performance of the mass spectrometer. This was and is necessary for routine automated collection and reduction of combined GC/HRMS data with minimal operator intervention in the procedures.

The system development is motivated by two goals. First, the system must be robust in the sense that it continue to operate under a variety of changing conditions, including intermittent misbehavior of the mass spectrometer. This ensures that the system can recover from hardware or software error conditions to prevent fatal "crashes" of the system and resulting loss of data. Second, the system must automate the GC/HRMS task. The volume of data acquired in GC/HRMS experiments can be efficiently handled only when every spectrum can be acquired and reduced for final output by the system without manual intervention. We are successful in these goals because we have written the software to determine the actual performance of the mass spectrometer and to have subsequent calculations based on that measured performance, as opposed to some hypothetical ideal.

We are now providing routine GC/HRMS service on a limited basis as we improve the system. The time required for system development and testing will slowly diminish over the next year, leaving additional time for analysis of mixtures obtained in our own work and that of our collaborators. We have deferred implementation of the metastable system (see below) while the GC/HRMS development is continuing, although we have completed the hardware and much of the software for the system.

1.2 Hardware Acquisition and Development

We have, in the mass spectrometry laboratory, two high resolution mass spectrometers, the Varian-MAT 711, and the AEI MS-9. Development efforts have focussed upon the MAT-711 because this more modern instrument is equipped with the high performance gas chromatograph needed for the GC/MS efforts.

We concurred with the study section's recommendation that stand-alone computer support be provided for efficiency and long-term cost effectiveness, and that such support be provided by the existing PDP 11/20 and a new PDP 11/45 or equivalent. We were able to adjust our first year budget to allow purchase of this computer.

At the time that the processor was ordered the cost of DEC disk drives was nearly double that of other vendor's drives. Accordingly we originally procured dual density top loading drives from System Industries. These drives were not directly software compatible with DEC RK type drives, but System Industries promised a hardware development to develop such compatibility and furnished software patches for DOS 8 so that we could use the drives. Unfortunately the hardware development was not carried out and our software needs expanded beyond DOS 8. We dealt with this problem by returning the System Industries' drive in the spring of 1975 and obtaining an equivalent drive for less money from International Memory Systems (IMS) which was RK compatible. The IMS disk drives have been installed for over a year with no indication of incompatibility with the DEC drives. The current hardware configuration is shown in Figure 1.

The PDP 11/20 processor is directly connected to the mass spectrometers and the gas chromatograph through two interfaces. The Ion Multiplier inter-face is a DR 11-B which provides direct memory access transfer of digitized ion multiplier samples. The direct memory access is necessary to provide a channel of sufficient bandwidth to achieve the requisite sampling rate for GC/HRMS work. The General Interface is our own design for a multiplexed interface used to select between the spectrometers; to manipulate the hardware mass scanner; to control the source voltage, the magnet current, or the analyser voltage; and to read the magnetic field, the source voltage, or the total ion current. All of these functions are slow speed and hence do not require the high data rate of a DMA interface. The general interface has 5 unused channels which are available for future development.

In addition to the instrument interfaces the PDP 11/20 is equipped with 8k of core memory, a KSR-33 terminal, a KW 11/P programmable clock and is tied to the PDP 11/45 via the Inter-Processor Interface (IPI). The IPI is a full duplex single word channel for which we have written a software driver providing user programs with 16 priority driven block transfer unidirectional channels. Thus, though the hardware provides only a single real channel it has proven easy to build mechanisms to

provide a very flexible and convenient mode of communication between the two processors.

The PDP 11/45 is equipped with 28k core, a PC 11 high speed paper tape reader/punch, an LA 30 terminal, a TM 11 industry compatible magnetic tape drive, an LP 11 300 line per minute printer, a KW 11/L line clock, a Loma Linda crt display, a CalComp drum plotter, and a hard line to the PDP 10. The dual Loma Linda / CalComp facility provides for both high speed real-time displays as well as for low speed off-line hardcopy graphics. The TM 11 provides a communications media to other processors and is used by procedures to save data on system failures and to maintain the archival data base.

1.3 Software Development

Conversion of existing PL/ACME programs to FORTRAN was begun on the award of the grant. Conversion of these algorithms also included many system software developments to ensure that previously batch processing programs could function in a real-time environment under the requirements of GC/HRMS operation. This development included not only improvements and extensions to existing algorithms, but building a file management system for facile logging and storage of spectra with the ability for simple recall to examine or recompute old data, and a diverse package of debugging, display and plotting and mass spectrometer evaluation programs. Development of improved capabilities for these tasks is an on-going project.

Because we view GC/HRMS as the most important new capability of our mass spectrometer/computer work, the requirements of GC/HRMS have guided development of the software system. These requirements include continuous automatic monitoring of instrument performance to avoid wasting time collecting poor or erroneous data. Because we have chosen to approach GC/HRMS with an electrical recording system, as opposed to photographic, we are able to monitor the instrument continuously, both during initial setup and during the course of the GC/HRMS experiment. Major sections of the software and how they interact among one another are summarized below.

During the past year the routine production usage of the HRMS data has become a reality. The direct utilization of the system for the acquisition of high resolution mass spectrometry data typically consumes 6 hours per day. This figure does not include time for the post-processing of data, retrieval of data from the archival data base, or for the generation of duplicate print outs of selected data. These demands add 1 to 2 hours of system service each day to the total high resolution system requirements.

Low resolution mass spectral data whether it be derived

from high resolution data or obtained directly as low resolution data, places additional time demands upon the data system. High to low resolution conversion, low resolution plotting, and low resolution spectral library searching have all generated a need for increasing amounts of system time.

In an effort to utilize the data system more completely during non-prime time, batch and spooling mechanisms have been constructed. The high resolution spectral reviewing mechanism may be actuated and then left unattended while the hard-copies are being generated. The high to low resolution conversion process contains a mechanism for the generation of a low resolution plotting spool which can be played without operator intervention. Batch procedures have been written which provide for the archival of newly acquired spectral data in the archival data base.

As with any system the size of the high resolution system there is a continual need for system maintenance and minor software upgrades. As a wider range of data acquisition and analysis becomes available new demands upon the system have developed which require modification of the software.

The net result of the production demands has been to reduce the amount of system time available for the development of new software facilities. Software development and production compete for the available system time reducing the productivity of both the chemical user and the software developer. This competition can be drastically reduced if software development can proceed on a machine separate from that on which production is done. The SUMEX PDP-10 offers an exceptionally attractive environment for software development. The TENEX operating system provides a more tractable medium for development than does the restricted environment provided by PDP-11 operating systems.

A major factor in the ease with which programs can be constructed is the ease with which text can be manipulated. The TV-EDIT program which is available on the PDP-10 has proven to be effective for this task. This program provides an extremely flexible text editing system for display terminals. The mechanics of program construction can be greatly simplified by the utilization of this facility. Typically all major (more than a few changes) text modification of programs are carried out on the PDP-10 using TV-EDIT and then transferred to the PDP-11. Thus even the task of writing FORTRAN programs is simplified even though there exist FORTRAN incompatibilities between the two machines.

While TV-EDIT has reduced development demands on the PDP-11 by eliminating PDP-11 text editing sessions, the problem of program compilation and debugging remain. Clark Wilcox, of the SUMEX staff, has provided an effective solution to this problem with the development of the MAINSAIL (machine independent SAIL) compiler. This compiler provides the user with a powerful,

machine independent, structured language. Not only is the compiler machine independent, but exhibits superior execution speeds and storage requirements as compared to the DOS 9 FORTRAN which has been used previously.

The combination of TV-EDIT and MAINSAIL has proven to be an effective method for the development of software for the PDP-11s within the PDP-10 environment. Most debugging can be carried out on the PDP-10 and then transferred to the PDP-11s for final debugging of machine-dependent facilities. The class of machine-dependent facilities includes device drivers and interaction with the operating system. The class of machine-independent facilities includes analysis algorithms, file manipulation, and most other programs which need development. This means that the amount of time required on the PDP-11 for program development can be reduced significantly using the aforementioned process, leaving more time for production demands.

1.4 Operating System

DOS version 8 was the first operating system to be used. However, this system was abandoned in favor of DOS 9. The major mandate for this conversion is the vastly improved overlay system offered by DOS 9. Overlaid files are maintained as a single, contiguous file on disk as opposed to the DOS 8 method of maintaining a separate linked file for each overlay. The DOS 8 strategy demands that a linked file be opened, read, and closed for each overlay load. DOS 9 allows an overlay to be loaded with a single disk read. Also the DOS 9 overlay facility provides for a tree structuring process which was completely absent from DOS 8. Considering that the version of the system in use at the time of the conversion had 17 overlays, the importance of efficient overlay loading is obvious. In addition to these factors, DOS 9 provides batch processing facilities which make it much easier to do system generation, archive data, etc.

We have been using DOS 9 for the past year. This operating system was chosen as the most suitable system available at the time we started its usage. Unfortunately DOS has many shortcomings. Bugs in many of the system programs and poor recovery from hardware errors on mass storage devices are the most visible defects. More subtle defects exist however when complex real-time processing is desired. These defects are compounded by our lack of monitor or system program sources.

In response to these defects in DOS we initiated an investigation into alternative operating systems. Both RSX-11M and RT-11 were examined in light of our particular demands. RSX-11M was rejected due to its size, poor terminal handling, and its implied dependence upon memory management hardware. RT-11 version 2C has been shown to possess advantages over both DOS and RSX-11M. RT-11 is a small system which comes as either a single

job monitor (1.5k word resident monitor) or a foreground/background monitor (3.5k word resident monitor). This is much smaller than RSX-11M (6k word resident monitor) and somewhat smaller than DOS 9 (about 4k word resident monitor). The foreground/background facility provides a convenient environment for simultaneous processing of plot, print, or filing spools with system program or user program execution. The single job monitor provides a small high speed system suited to real-time instrument control and data acquisition.

Both the I/O facilities and file structure of RT-11 possess advantages over those provided by DOS. RT-11 provides a queue structure for all I/O, leading to a more flexible utilization of peripherals. Additionally a completion routine facility is available which allow user supplied routines to be invoked upon I/O completion, providing interrupt service outside of the device drivers. Adding, deleting, or modifying a device driver is also very easy, amounting to simply replacing a file on the system device. While the file structure is limited to contiguous files the access time to these files is much more rapid than that provided by DOS. The rapid file access is quite evident when running system programs. Assembly, linking, and file transfer operations are significantly faster operations under RT 11 than under DOS 9. This is an important consideration in light of the fact that it takes over 35 minutes to link the GC/HRMS system under DOS 9 and such slow response seriously degrades programmer efficiency.

RT-11 is additionally attractive in light of the development of MAINSAIL. The runtime system for MAINSAIL under RT-11 already exists while none is available for DOS. The RT-11 magnetic tape formats are directly readable and writeable by the PDP-10, eliminating the conversion necessary for DOS magnetic tape files. The RT-11 system will also provide a much cleaner interface for the hardline to the PDP-10. It will be possible to log onto the PDP-10 through the PDP-11 RT-11 system and transfer files directly between the systems via the hardline.

1.5 Combined Gas Chromatography/High Resolution Mass Spectrometry

The gas chromatography/high resolution mass spectrometry (GC/HRMS) system provides for the acquisition, analysis, and archival storage of high resolution mass spectral data of gas chromatographic effluents. The system is composed of a real-time instrument control and data acquisition system, a post-processing system, an archival data base, and various development facilities.

SAQMON is an assembly language real-time instrument control and data acquisition monitor which executes within the PDP 11/20. SAQMON is responsible for controlling and monitoring all

instrument hardware to provide for the acquisition of high resolution data from the mass spectrometer. It contains processes to start and stop mass scanning in both a cyclic and single scan fashion. DC signal level is determined here and peak thresholding and background removal are also done here. A major portion of the memory allocated to SAQMON is dedicated to buffering of the peak profile data, relieving the PDP 11/45 processor of this burden.

SAQMON communicates with the PDP 11/45 through the Interprocessor Interface using the IPIDVR program which provides 16 unidirectional priority driven channels between the processors using the IPI. Such a scheme allows for independent communication between the systems depending on the task being performed and the data being acquired.

REFRUN is a FORTRAN overlayed program which is responsible for the acquisition, filing and post-processing of high resolution calibration spectra. Prior to analyzing a sample of interest the instrument must be calibrated by generating spectra of a reference gas (currently perfluorokerosene) which can be later used to compute the masses of ions acquired in spectra of unknowns. REFRUN uses SAQMON to acquire peak profile (PPF) data from the instrument or the IOLNK program to acquire PPF data from a back up file. PPF data is converted to mass/amplitude pairs and various characteristics of the spectrum are computed. These results are summarized in a CRT display for use by the operator. This summary includes the calibration range, the voltage of the reference base peak, and plots of a model peak, the projection error versus mass and the resolution versus mass. From this summary the operator can gauge the performance of the total system. The model peak plot provides critical information on the instrument set-up so that the operator can optimize the instrument performance. Once the operator can repetitively calibrate using the reference gas a spectrum is filed. Both the PPF and the reduced data are filed so that all system functions can be performed again at a later time. When the data is filed automatic displays are generated of the scan summary and mass/amplitude pairs. REFRUN also provides a reviewing capability so that reduced data files can be used to generate additional copies of the displays.

SAMRUN is an overlayed FORTRAN program which executes within the PDP 11/45 to acquire, analyze, and post-process spectra of samples. SAMRUN uses SAQMON to acquire PPF data from the instrument or the IOLNK to acquire PPF data from a backup file. Spectral analysis of samples requires a reference spectrum previously filed by the REFRUN program. The reference spectrum is used to guide the detection of reference peaks within the spectrum of the sample. The reference peaks which are found in this fashion provide discrete samples giving the time-mass conversion information. Mass values are computed by interpolation between the reference peaks. The spectrum summary presented to the operator for each sample spectrum is similar to

that provided by REFRUN minus the graphics plus information on the amplitude of the sample's base peak. When a set of sample spectra are filed both the PPF and the reduced data are filed, providing the same rerun facilities as REFRUN. The automatically generated displays include the spectra summaries, the mass/amplitude listings and the composition listings. SAMRUN also provides a reviewing capability for generating new copies of displays or new composition listings with different parameters.

The minute quantities of certain samples which have been submitted for analysis prohibit the re-running of any experiments associated with these samples. The system operates in a somewhat hostile environment. The physical laboratory environment dictates that the computer system be located in close proximity to the GC/MS instrument. The instrument can cause severe electromagnetic disturbances (sparks within the source, high voltage shut down, etc.) which can bring down either the entire data system or portions of the system. Static electric discharges from the operator through the system console have also resulted in catastrophic consequences for the data system. These occurrences are quite unpredictable from the software point of view and are difficult to alleviate in the physical environment. Therefore, the software must file data as soon as it is acquired in order that in the event of system failure any data gathered up to that point is maintained intact. A restart facility is also provided so that an experiment can be continued after catastrophic failure, losing only the data associated with the particular mass scan in progress at the time of the failure.

Both raw and reduced data are logged in real-time into a standard system file. The operator has the option of permanently filing this data in a file with an automatically generated name or to ignore the experiment altogether and file none of the data. Filing of both the raw and reduced data is necessary so that later rerunning of the experiment can be carried out. This is desirable in case of difficult data or in cases of software malfunction.

Buffering is a central issue in the system. Due to the uneven distribution of data, high data rates, and slack periods, it is desirable to provide a large amount of buffering between the instrument itself and the reduction processes. It is the case that data from one spectrum can be reduced while another spectrum is being acquired. Currently the PDP 11/20 has sufficient buffer capacity to hold almost a complete spectrum. The PDP 11/45 can concentrate on the conversion of time/intensity information into mass/amplitude information and the generation of displays with little regard to buffering the raw data.

Feedback provided by the real-time displays can be used by the operator to determine the quality of the spectral data. One can disregard scans which are poor and know when one is of high quality. The operator can choose to print out results immediately for critical samples, or defer final output until

later while additional data are being collected. An archival system provides the facility for storing and retrieving old spectral data for review or reanalysis.

High resolution mass spectral data often contain peak complexes consisting of more than one peak not separated by the simple thresholding technique. This problem is aggravated in GC/HRMS experiments because scans are acquired at lower resolving powers to achieve increased sensitivity. In GC operation, a further source of overlapping peak complexes is bleed from the organic phase of the GC column; many components of column bleed have masses similar to those of perfluorokerosene, the reference material. In particular if a bleed peak is so close to a reference gas peak used for calibration that a complex arises the entire calibration mechanism can go awry. In response to this problem we have developed a technique for the analytic resolution of such complexes. The problem has two aspects. First, a reliable detection method for complexes must be available. The computer must be able to tell the difference between single peaks and complexes of peaks. Second, once the computer detects a complex it must be able to provide an estimate of the position and area of the component peaks. After careful examination of the data it was determined that a reliable detection technique could be based upon the second moment of peaks suspected of being complexes. The basic idea is to determine the statistics of a peak which is representative of a single peak in the (mass) region of the suspected complex. A decision can then be based upon a comparison of the observed 2nd moment and the 2nd moment of the representative peak. It should be noted that the representative peak is dynamic within each scan due to instrumental variations in the resolution vs. mass curve. Once a complex is detected it is subjected to an analytic resolution technique developed by our personnel which computes the position and area of two peaks assumed to produce the complex. This technique works on the previously calculated statistics of the representative peak and the actual statistics of the observed complex. This method of resolving peak complexes has made possible the full reduction of GC/HRMS data which is not reducible otherwise. The operator has the option of either normal data reduction or data reduction with the resolution technique.

1.6 High Resolution Spectra Utility Programs

The development of the GC/HRMS system has generated some additional programs for the examination of peak profile data. These programs are not intended for usage by the chemists but rather serve as tools for the software personnel developing new facilities and analyzing failures of existing facilities. PPFSEE is a FORTRAN program which allows the user to plot on a CRT or CALCOMP the profiles of selected peaks from a spectrum. The relevant statistics of the peak area, amplitude, width, 1st, 2nd,

and 3rd moments are also displayed. This program is useful for examining peaks for the occurrence of doublets and comparing peak shapes obtained under differing instrument conditions. PKEXAM is a FORTRAN program which provides the user with various spectral plots. 2nd moment vs time is a typical output which is used to evaluate the performance of the peak complex resolution mechanism.

Often the investigator submitting a sample for GC/HRMS GCMS can obtain useful information from a low resolution plot of the high resolution data. HRTOLR is a FORTRAN program which converts reduced high resolution data to a standard low resolution format. This conversion can be carried out in two modes:

- 1) All peaks which are present in the sample spectra but not in the reference spectra are represented in the low resolution output.

- 2) Only peaks whose masses match a user supplied composition are represented in the low resolution output.

Available in both of these modes are facilities for PFK removal, selected mass removal, scan selection, compound renaming, and spooling for later low resolution plotting.

We currently support two types of low resolution data post-processing. First, the program LRLOT produces plots of low resolution spectra. It is capable of generating plots of individual spectra of a selected file or plotting of all files contained in a spool file which can be produce either by the HRTOLR program or with a text editor. Secondly, the program SEARCH (developed by ourselves and our collaborators in the department of Genetics) can be used to search a library of low resolution spectra for matches to a user supplied spectrum. Thus data acquired from either high or low resolution operation can be plotted and library searched.

1.7 Process Monitor: PMON

Experience has shown that it is very difficult to obtain full processor utilization with a traditional subroutine structure. Such a structure lends itself to predefined static conditions rather than to the dynamic situation presented by a real-time instrument. The operator requires automated functions to be available in unpredictable ways due to the experimental nature of the work being done. What is required is a method for scheduling program execution on a priority demand basis.

PMON is a monitor for scheduling real-time processes in a user definable fashion. Each system which uses PMON simply links PMON as the main segment of the program. The user supplies a process structure, PSTRUC, which describes to PMON all processes

in the system and their relative priorities . The PSTRUC contains a sequential list of priority levels running from the highest priority through the lowest priority. Each priority level is composed of a ring of process descriptors which specify processes at the same level. All processes represented on a given level are guaranteed to receive equal processor attention. Each process has associated with it a list of blocking conditions. These conditions are simply booleans relating to the state (empty/non-empty) of a queue. PMON schedules the highest priority process that has a satisfied blocking condition. Processes communicate with each other by pushing information into queues and by waiting for queues to attain a desired state. The queues are manipulated in a mutually exclusive fashion so that completion routines can send information to processes about I/O completion at an interrupt level. The current implementation of PMON requires less than 512 words of memory, despite its implementation in a high level machine independent language.

1.8 METASYS

METASYS is a data acquisition and analysis system for data on metastable ions. It is constructed around the PMON real-time monitor. It is composed of two autonomous subsystems:

- 1) A High Resolution Metastable Virtual Instrument (MVI)
- 2) A Data Manipulation System (DMS)

The MVI provides the user with the following capabilities:

- 1) Setting of any automated control.
- 2) Reading any automated indicator.
- 3) Scan source voltage, analyser voltage, and magnet current in user selectable fashions.
- 4) Acquire digitized samples of ion multiplier current, magnetic field, source voltage, total ion current, or functions of these samples in a user selectable fashion.
- 5) The generation of the Context Base which is a medium term memory for system events. All operator interaction and all data acquired from the instrument are recorded here.

The DMS provides the user with the following capabilities:

- 1) Permanent filing of data contained in the Context Base into the METASYS Data Base. (MDB).
- 2) Reexamination of data contained within the MDB.